Faculty of Engineering and Information Technology
University of Technology, Sydney

Contrast Mining in Large Class Imbalance Data

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Jinjiu Li

May 2013

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

_____

Acknowledgments

First and foremost, I would like to express the deepest appreciation to my supervisor, Professor Longbing Cao, for providing me with all the necessary facilities and his great effort in leading me to the spirit of freedom in regard to research. Without his guidance and persistent help this thesis would not have been possible.

Thanks to my co-workers, Can Wang and Wei Wei, for being so supportive especially when I was struggling through hard times.

I am also grateful to my dear friends and colleagues in AAI, especially fellow students Mu li and Chunming Liu for their hard working in my project teams.

I also place on record, my sense of gratitude to the team members in Westpac for their expert and sincere help in the project.

I would like to express my eternal appreciation towards my family, my wife, my daughter Jessie and my son Vincent. Without their unconditional love, I could not have outlasted those tough times.

Jinjiu Li
December 2012 @ UTS

Contents

List of Figures

List of Tables

Abstract

Class imbalance data, in which the classes are not equally represented and the minority classes include a much smaller number of examples than other classes, is pervasive and ubiquitous, particularly in applications such as fraud/intrusion detection, medical diagnosis/monitoring, and risk management. The conventional classifiers tend to be overwhelmed by the large classes while ignoring the smaller classes. Typically, many of the existing solutions to the class imbalance problem are proposed at the data level, and a few at the algorithmic level. However, the prior methods have more or less limitations in anomaly detection according to our extensive experiments. Therefore, the thesis targets contrast mining to solve the problem of anomaly detection in imbalanced data from three aspects: feature construction, an effective algorithm for mining contrast patterns, and selection of optimal rule combinations through analysing rule interactions.

Feature construction is one of the most important steps in contrast pattern mining, and any other data mining processes as well. The majority of feature construction methods, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Fourier Transformation, and Independent Component Analysis, usually generate new features by transforming the existing raw features into a new data space. Therefore, previous solutions have many limitations with respect to the objective of training highly accurate classifiers in class imbalance data sets. Incomprehensible features may be generated, based on the assumption that all the samples are independent, the feature set is unstable and sensitive to trivial change of the sample set,

it is difficult to integrate significant domain knowledge, and the classifiers built on the transformed feature set suffer from high False Positive Rate in the class imbalance data set.

In order to train high performance models in the imbalance scenario, we propose a novel method, Personalised Domain Driven Feature Mining (*PDDFM*), to generate important features by integrating domain knowledge effectively with a full consideration of the correlations among samples. A framework specially designed for *PDDFM* is introduced. A novel feature selection method, called Mutual Reduction, is proposed to minimise the noise from redundant features and maximize the contribution of "trivial" features whose gain ratio are low but contribute positively when cooperate with the others. The experimental evaluation reveals our feature mining approach outperforms state-of-the-art methods in anomaly detection.

Contrast pattern mining has been studied intensively for its strong discriminative capability. However, state-of-the-art methods rarely consider the class imbalance problem, which has been proven to be a significant challenge in mining large scale data. The thesis introduces a novel pattern, i.e. converging pattern, which refers to the item sets whose supports contrast sharply from the minority class to the majority class. A novel algorithm, ConvergMiner, is also proposed to mine converging patterns efficiently. A light-weighted index T*-tree is built to speed up the search process, and output patterns instantly. A series of branch bound pruning strategies are further presented to greatly reduce the computational cost. Substantial experiments on large scale real-life online banking transactions for fraud detection show that the ConvergMiner greatly outperforms the existing cost-sensitive classification methods in terms of accuracy. In particular, it efficiently and effectively detects the frauds in large-scale imbalanced transaction sets. More importantly, the efficiency improves with the increase in data imbalance. After many converging patterns are generated, we propose an effective novel method to select the optimal pattern set.

Rule-based anomaly and fraud detection systems often suffer from sub-

stantial false alerts in the context of a very large number of enterprise transactions with class imbalance characteristics. A crucial and challenging problem is to effectively select a globally optimal rule set which can capture very rare anomalies dispersed in large-scale background transactions. The existing rule selection methods which suffer significantly from complex rule interactions and overlapping in large imbalanced data, often lead to very high false positive rates. We analyse the interactions and relationships between rules and their coverage in transactions, and propose a novel metric, *Max Coverage Gain* (*MCG*). *MCG* selects the optimal rule set by evaluating the contribution of each rule in terms of overall performance to cut out those locally significant, but globally redundant rules, without any negative impact on the recall. An effective algorithm, *MCGminer*, is then designed with a series of built-in mechanisms and pruning strategies to handle complex rule interactions and reduce computational complexity in identifying the globally optimal rule set. Substantial experiments on 13 UCI data sets and a real time online banking transactional database demonstrate that *MCGminer* achieves significant improvement in accuracy, scalability, stability and efficiency with respect to large imbalanced data compared to several state-of-the-art rule selection techniques.

Following that, the above proposed contrast analysis techniques have been applied in two industrial projects. The first project was "Fraud Detection in Online Banking" for a major bank in Australia. We developed a risk management platform called *i-Alertor* , which is mainly powered by the techniques introduced in this thesis. According to the evaluation report, i-Alertor outperforms the existing rule based system by 10%. The second project was the "Key Indicator Discovery in Student Learning" for a key University in Australia. Another platform called i-Educator is also developed to support this application.